A SIMPLE FRAMEWORK FOR LOW-RESOLUTION DETECTION WITH HIGH-RESOLUTION KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper dedicates to improving object detection performance on low-resolution images. The intuitive way is to distill the high-resolution knowledge from models trained over high-resolution images, shorted as cross-resolution distillation. Unfortunately, most of existing conventional distillation methods focus on the knowledge distillation with same-resolution images in both teacher and student. Directly applying these methods for the cross-resolution distillation results in limited improvement. To address this issue, we introduce a simple yet effective framework, *i.e.*, LRDet. The key in LRDet is the *bridge branch*, acting as an intermediate status between teacher and student. With the bridge branch, LRDet can i) align the resolution and supervision targets between the high-resolution teacher and the low-resolution student, and ii) then transfer the high-resolution knowledge smoothly and effectively. Experiments demonstrate that LRDet consistently improves various well-known detectors on low-resolution images, *e.g.,* from 35.4 mAP to 37.8 mAP with RetinaNet-R50 on MS COCO using 600×1000 input. Meanwhile, it is easy to utilize large teachers in LRDet as the conventional distillation methods do, which can further improve the low-resolution performance. For example, RetinaNet-R50 with 600×1000 resolution can achieve 39.7 mAP when distilling from RetinaNet-X101.

1 INTRODUCTION

High-resolution (HR) images are essential for position-sensitive tasks such as object detection Tan & Le (2019). Current detectors trained over HR images achieve top performance on various benchmarks He et al. (2016); Lin et al. (2017b); Tian et al. (2019b); Ren et al. (2015); He et al. (2017); Wang et al. (2020a); Lin et al. (2014). But HR images increase model computational costs, slow down the model inference speed Redmon et al. (2016); Redmon & Farhadi (2018)[1], and may not be available in specific scenarios Kim et al. (2016). Differently, the easily collected low-resolution images lead to smaller computational costs and faster speeds. While low-resolution images may lose fine visual details due to the low image quality, which severely degrades the performance. In this paper, we aim to improve object detection on low-resolution images.

An intuitive way to achieve the above goal is to use knowledge distillation methods to transfer high-resolution knowledge from the teacher model trained over high-resolution images (HR teacher) to the student model with low-resolution images (LR student). Existing conventional distillation frameworks Heo et al. (2019); Yim et al. (2017); Zagoruyko & Komodakis (2016); Yang et al. (2022) are proposed to transfer knowledge from teacher to student with the same resolutions. Directly applying these distillation methods, including the state-of-the-art FGD Yang et al. (2022), to the cross-resolution setting results in limited gains (Table 1). It shows that conducting knowledge distillation among different resolutions is a non-trivial task. Although the resolution gap between the feature maps can be easily eliminated by feature interpolation (upsampling for students or downsampling for teachers), the mismatch of the supervision targets between the HR teacher and the LR student still exist. This mismatch makes it difficult to utilize high-resolution knowledge to help the training of the LR student.

[1]Model computational cost is quadratic to the image resolution. High-resolution images bring significant computational overhead and slow down model inference speed.

Table 1: Cross-resolution distillation with LRDet *vs.*various conventional methods. We provide the baseline model (RetinaNet-R50 Lin et al. (2017b) trained over 600×1000 resolution) in the second row. All other experiments are conducted on MS COCO dataset Lin et al. (2014) with RetinaNet-R50 trained over 800×1333 resolution as the HR teacher and 600×1000 resolution as the LR student. Up and Bridge represent upsampling and bridge branch, respectively.

Method	Align	AP	AP_S	AP_M	AP_L
–	–	35.4	18.0	39.5	48.3
Fitnet Romero et al. (2014)	Up	36.0	19.1	40.2	48.6
CWD Shu et al. (2021)		36.1	19.0	40.3	48.4
DeFeat Guo et al. (2021)		36.2	19.3	40.5	48.7
FGD Yang et al. (2022)		36.3	19.2	40.7	48.6
LRDet	Bridge	37.1	20.1	40.8	50.1

Given the above analyses and quantitative experiments in Table 1, we present LRDet as a novel framework to facilitate feature distillation between the HR teacher and the LR student. The key insight in LRDet is the proposed bridge branch. Specifically, the bridge branch consists of FPN and detection head, takes high-resolution feature maps (upsampled from outputs of the LR student backbone) as input, and is supervised by ground-truth boxes in high resolutions. The design of the bridge branch matches its inputs and targets with the HR teacher. Meanwhile, since there is no resolution gap between the bridge branch and the HR teacher, the high-resolution knowledge can be smoothly transferred as in previous distillation frameworks. Moreover, by sharing parameters, the LR student perceives high-resolution knowledge from the bridge branch, which makes the knowledge distillation across resolutions can be conducted effectively. Compared with the conventional frameworks, LRDet achieves knowledge distillation across resolutions in a simple and effective way. In addition, we find that weight inheriting strategy is another key to improve the low-resolution detection, which initializes the weights of the LR student with the HR teacher.

We conduct extensive experiments on MS COCOLin et al. (2014) to show the effectiveness of our LRDet with different low resolutions (600×1000 and 400×667). Results show that LRDet is compatible with various detectors (RetinaNet Lin et al. (2017b), FCOS Tian et al. (2019b), and Faster R-CNN Ren et al. (2015)) and achieves significant improvements. Sometimes the LR models trained with LRDet even surpass their HR teachers (800×1333) while requiring much less computational cost. Further, we validate the generalization of LRDet on two other tasks, *i.e.*, instance segmentation and human keypoint detection. LRDet consistently improves the performance of the low-resolution models, which indicates that our framework has strong generalization to other position-sensitive tasks.

2 Related Works

Object detection is a fundamental and challenging task in computer vision, which involves classifying individual objects and localizing each using a bounding box Girshick et al. (2014). The CNN-based detection models are divided into two-stage Ren et al. (2015); He et al. (2017); Cai & Vasconcelos (2018) and one-stage detectors Lin et al. (2017b); Tian et al. (2019b); Yang et al. (2019); Duan et al. (2019); Ge et al. (2021). Faster-RCNN Ren et al. (2015) is a typical two-stage detector that utilizes RPN to achieve high-performance detection. RetinaNet Lin et al. (2017b) as a one-stage detector is able to perform detection faster, and it combines FPN Lin et al. (2017a) and FCN Long et al. (2015) to make dense detection on feature maps directly. However, one-stage detection relies on a large number of prior anchor boxes, which brings extra computation. Anchor-free detectors, e.g., FCOS Tian et al. (2019b), alleviate this problem. They predict the key points and location of objects instead. The design of these detectors generally focuses on feature representation or model structure rather than on the input resolution. In this paper, we focus on improving the detection performance with low-resolution images.

Knowledge distillation is a method of knowledge transferring and model compression. It is first proposed to distill the knowledge from a large teacher model to a compact student model for the classification task Yim et al. (2017). Over the years, many improved KD methods have been proposed that perform distillation over intermediate features Romero et al. (2014); Tian et al. (2019a), relation representation Park et al. (2019); Tung & Mori (2019), attention Zagoruyko & Komodakis (2016),

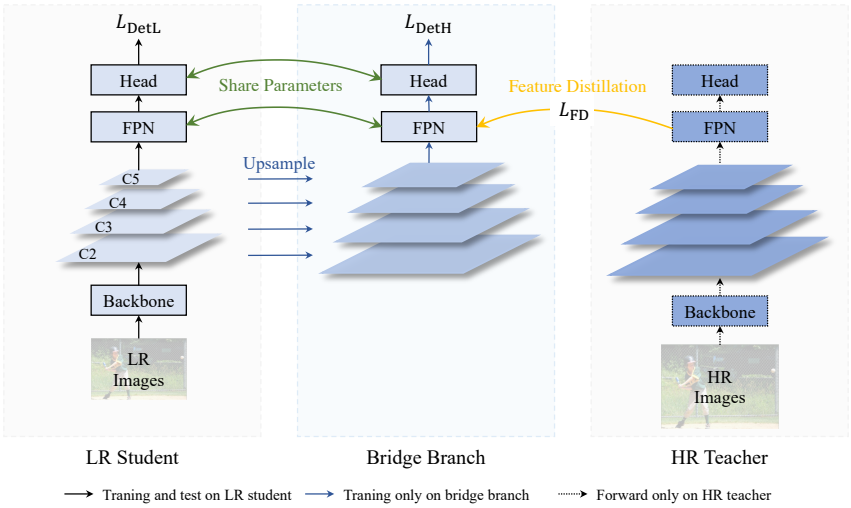

Figure 1: **LRDet** framework. LRDet aims to improve the performance of object detection on low-resolution images, *i.e.*, LR student in the figure. The key in LRDet is the bridge branch, which can learn high-resolution knowledge from the HR teacher based on the feature distillation and transfer the learned information to the LR student. The inputs of the bridge branch are produced by upsampling LR features from the LR student backbone. Best viewed in color.

etc. Recently, some works have successfully applied KD to detection Chen et al. (2017); Li et al. (2017); Wang et al. (2019); Zhang & Ma (2020); Guo et al. (2021); Dai et al. (2021); Yang et al. (2022). FGD Yang et al. (2022) is a powerful feature-based distillation method. It decouples the foreground and background of the image and uses focal and global distillation to guide the student model, achieving remarkable results. However, these efforts mainly focus on feature transfer at the same resolution and do not consider the distillation across resolutions, which makes them work ineffective with a resolution gap. We consider distillation with different resolutions and design a bridge branch to help the knowledge transfer between the HR teacher and the LR student.

There are also previous works that introduce auxiliary branches to facilitate the knowledge distillation, such as TAKD Mirzadeh et al. (2020) in image classification. TAKD aims to bridge the capacity gap and introduces an independent assistant network. It conducts distillations individually between teacher and assistant, assistant and student. Differently, our LRDet takes efforts to resolve the mismatches between the HR teacher and LR student. Our bridge branch shares knowledge with the LR student and there is no distillation between them.

Improving low-resolution detection is challenging. Multi-scale training is a common technique to enhance the model robustness against resolution variation He et al. (2015b); Singh & Davis (2018); Singh et al. (2018); Wang et al. (2020b), so it can be an effective approach to enhance low-resolution detection. Multi-scale aligned distillation (MSAD) Qi et al. (2021) combines multi-scale training and KD to improve the low-resolution students. It proposes a multi-scale fusion network that serves as the bridge. In the MSAD framework, LR student adopt larger feature maps to solve the size gap, *i.e.*, it uses P2$\sim$P6 level features for LR detection on FCOS Tian et al. (2019b) instead of standard P3$\sim$P7 level. And the proposed fusion network narrows the semantic gap. However, their framework has constraints on resolution and are computationally costly. In this paper, we retain the advantages of low resolution inputs, *i.e.*, low cost and fast speed, and propose a simple but effective method to improve the low-resolution detection via feature distillation.

3 LRDET

The overview of LRDet framework is shown in Figure. 1. The main component in LRDet is the *bridge branch*, acting to release the non-trivial problem when transferring knowledge from the HR teacher to the LR student. Under the help of bridge branch, LRDet can effectively transfer the

rich knowledge from the HR teacher to the LR student with the feature distillation optimization. Moreover, the weight inheriting strategy is also effective in the LRDet.

3.1 BRIDGE BRANCH

As analyzed in the previous section, the supervision mismatch between the HR teacher and the LR student brings knowledge gap that hinders the cross-resolution distillation. We thus design a bridge branch for supervision alignment between the LR student teacher and the student, to smoothly and effectively transfer knowledge between resolutions.

To resolve the supervision mismatch, we extend the LR student by adding a branch for detection on high-resolution feature maps, in parallel with the student's detection on low-resolution features. Specifically, the bridge branch consists of the FPN and the detection head, sharing weights with the same layers in the LR student. The input of the bridge branch is the HR features, produced by upsampling the LR features from the LR student backbone. These HR features are then fed into the bridge branch to obtain the FPN pyramidal HR features and the HR detection predictions, which are supervised by respective loss functions (refer to next).

The designed bridge branch has two characteristics. One is that the features from the bridge branch have the same resolution with those from the HR teacher, eliminating the resolution gap between the HR teacher and the LR student. More importantly, the bridge branch performs high-resolution detection with the same parameters of the LR student, which aligns the supervision between teacher and student. Therefore, the supervision mismatch is resolved and the high-resolution knowledge can be smoothly transferred.

Optimization. With the bridge branch, the high-resolution knowledge can be smoothly transferred from the HR teacher to the LR student. For the optimization of the bridge branch, we formulate the loss function L_{Bridge} as follows:

$$L_{\mathrm{Bridge}} = L_{\mathrm{DetH}} + \alpha L_{\mathrm{FD}}, \tag{1}$$

where L_{DetH} denotes the standard detection loss on the HR predictions from the bridge branch. L_{FD} represents the feature distillation from the FPN pyramidal features in the HR teacher to those in the bridge branch. Here, we simply use mean squared error (MSE) loss as the feature distillation loss. α is the loss weight hyperparameter.

Discussion. *Can feature distillation be replaced by other distillation algorithms?* Sure. Our LRDet can compatible with various distillation algorithms. In this paper, we mainly focus on the design of the bridge branch and the simple feature distillation already can achieve high performance.

3.2 WEIGHT INHERITING

A good weight initialization is beneficial to the model optimization Glorot & Bengio (2010); He et al. (2015a). It motivates us to carefully consider the initialization of the LR student. Inspired by the inherit initialization strategy Yang et al. (2022); Kang et al. (2021), we initialize the weights of the LR student with those in the same layers from the HR teacher. In this way, the student model and the bridge branch have strong feature representation at the beginning of the training. This weight inheriting strategy has merits of: i) making the LR student easy to converge, and ii) making the bridge branch and the student more receptive to the knowledge from the HR teacher.

3.3 OVERALL

The whole training process of LRDet framework is straightforward. First, the LR images are processed by the LR student backbone to produce the LR feature maps. These feature maps are normally passed to the later structure to obtain the LR predictions. Meanwhile, HR feature maps are upsampled from these LR feature maps and then fed into the the bridge branch, allowing the network to transfer high-resolution knowledge from HR teacher and generate the HR predictions. The overall training supervision:

$$L = L_{\mathrm{DetL}} + L_{\mathrm{Bridge}}, \tag{2}$$

where L_{DetL} is the traditional detection loss on the LR predictions, and L_{Bridge} is the loss function for the optimization on the bridge detector (see Eq. equation 1).

During the inference phase, only the LR student is enabled to make the final detection predictions with LR images input, while the HR teacher and the bridge branch are both discarded.

4 EXPERIMENTS

We evaluate our LRDet with a range of detectors (RetinaNet Lin et al. (2017b), FCOS Tian et al. (2019b), and Faster R-CNN Ren et al. (2015)) on the MS COCO benchmark Lin et al. (2014). We first show that LRDet brings significant improvements over the baselines when training with LR images. Then, we provide comprehensive ablation studies to show the effectiveness of each component of LRDet. For fair comparison, we use the MMdetection codebase Chen et al. (2019) and follow all the settings and hyper-parameters. Other settings are as follows.

Self-distillation Setting. In this section, most of our experiments are conducted in self-distillation setting, *i.e.*, the teacher and student models keep the same architecture (*e.g.*, using RetinaNet-R50 to distill RetinaNet-R50), and they differ only in the input resolution. The reason for this is to eliminate interference, if a larger teacher is adopted (*e.g.*, using RetinaNet-Res101 to distill RetinaNet-R50) as in the general KD approaches, we cannot distinguish whether the distillation gain comes from the high-resolution or the larger model capacity. However, we must point out that our LR student can be improved more significantly if not limited to self-distillation setting.

Resolution Settings. In the single-scale setting, we adopt 400×667 or 600×1000 low-resolution images. And we upsample these LR images to 800×1333 and use them as high-resolution images. In the multi-scale setting, the resolution ranges are $[320 \times 667, 400 \times 667]$ and $[480 \times 1000, 600 \times 1000]$ for two low resolutions while the high-resolution range is $[640 \times 1333, 800 \times 1333]$. Note that experiments in this paper only use on the above resolutions to verify the effectiveness of LRDet, but resolutions are not limited.

Loss Weight α. Loss Weight α is a hyper-parameter to balance the detection loss and the feature distillation loss in the bridge branch. For simplification, we set $\alpha = 0.005$ for all experiments on various detectors. We believe that if tuning α carefully, we can achieve higher performance.

4.1 MAIN RESULTS

In this section, we conduct experiments on a variety of detectors and provide comparisons between the models with or without applying LRDet when training with low-resolution images in the single-scale and multi-scale settings. Note that in all experiments in this section, the teachers and students keep the same networks, and their difference is only the resolution of the input image, which is different from the conventional knowledge distillation. For training details, HR teachers are trained with a $2\times$ schedule in the single-scale setting and with a $3\times$ schedule in the multi-scale setting.[2] For student models, a $1\times$ schedule is adopted. All experiments in this section are conducted with ResNet-50 as their backbones. Moreover, we provide extra results trained with a longer $3\times$ schedule in the multi-scale setting, proving the effectiveness of LRDet over the multi-scale training technique.

Comparisons in Single-scale Setting. Table 2 shows that LRDet improves RetinaNet Lin et al. (2017b), FCOS Tian et al. (2019b), and Faster R-CNN Ren et al. (2015) by 2.4 mAP, 1.6 mAP, and 2.1 mAP when training with a low resolution 600×1000, presenting even better results than the HR teachers. When the resolution decreases to 400×667, LRDet also shows significant gains. The above achievements are nontrivial as we adopt the same networks for teachers and students. In this setting, the previous state-of-the-art distillation method FGD Yang et al. (2022) only gives a 0.8 mAP on FCOS. Our designed bridge branch helps the LR student to obtain the size-robust features, and it further facilitate the knowledge transferring. Therefore, our LRDet can provides consistent gains across small, medium, and large objects.

Comparisons in Multi-scale Setting. Multi-scale training is a solid technique to improve detection performance and works well in low-resolution detection. We apply multi-scale training for Reti-

[2]All $1\times$, $2\times$, and $3\times$ schedules are standard schedules in MMdetection Chen et al. (2019), which represents training the models for 12 epochs, 24 epochs, and 36 epochs.

Table 2: Results on MS COCO in single-scale setting. For each detector, we first report the model performance with straightforward training at 800px, 600px, and 400px resolution. Our LRDet utilizes the above 800px models to transfer knowledge.

Detector	Resolution	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet-R50 Lin et al. (2017b)	800px	37.4	56.7	39.6	20.0	40.7	49.7
	600px	35.4	54.1	37.9	18.0	39.5	48.3
	400px	32.1	49.5	34.1	12.3	36.4	48.3
	600px(ours)	37.8	57.0	40.3	20.2	41.8	51.1
	△	(+2.4)	(+2.9)	(+2.4)	(+2.2)	(+2.3)	(+2.8)
	400px(ours)	33.9	51.3	36.4	14.2	39.1	49.9
	△	(+1.8)	(+1.8)	(+2.3)	(+1.9)	(+2.7)	(+1.6)
FCOS-R50 Tian et al. (2019b)	800px	38.7	57.4	41.8	22.9	42.5	50.1
	600px	37.1	55.5	39.9	19.8	41.1	50.2
	400px	33.7	51.2	35.9	15.2	37.3	49.8
	600px(ours)	38.7	57.1	41.6	21.8	42.7	51.5
	△	(+1.6)	(+1.6)	(+1.7)	(+2.0)	(+1.6)	(+1.3)
	400px(ours)	34.8	52.2	37.0	16.2	38.6	50.4
	△	(+1.1)	(+1.0)	(+1.1)	(+1.0)	(+1.3)	(+0.6)
FasterRCNN-R50 Ren et al. (2015)	800px	38.4	59.0	42.0	21.5	42.1	50.3
	600px	36.6	57.2	39.5	19.4	40.4	49.1
	400px	34.1	53.4	36.8	15.0	37.5	49.6
	600px(ours)	38.7	59.2	42.2	21.5	42.6	52.3
	△	(+2.1)	(+2.0)	(+2.7)	(+2.1)	(+2.2)	(+3.2)
	400px(ours)	36.0	55.4	38.8	16.9	39.9	51.2
	△	(+1.9)	(+2.0)	(+2.0)	(+1.9)	(+2.4)	(+1.6)

Table 3: Results on MS COCO in multi-scale setting. For each detector, we first report the performance at different resolutions with multi-scale training. Our LRDet adopts the 800px-mstrain-3× models as the teachers. Conventionally, we also report the results with 3× training schedule.

Detector	Resolution	Schedule	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet-R50 Lin et al. (2017b)	800px	3×	39.5	58.8	42.2	23.8	43.2	50.3
	600px	1×	35.5	54.2	37.7	19.0	39.5	47.9
		3×	38.3	57.3	40.8	21.0	42.6	51.0
	400px	1×	31.8	49.1	33.6	12.7	36.6	47.4
		3×	34.7	52.5	36.6	14.5	39.5	50.6
	600px(ours)	1×	38.6	58.0	41.0	21.7	42.8	51.5
		△	(+3.1)	(+3.8)	(+2.3)	(+2.7)	(+3.3)	(+3.6)
		3×	39.1	58.4	41.3	22.6	43.2	52.0
		△	(+0.8)	(+1.1)	(+0.5)	(+1.6)	(+0.6)	(+1.0)
	400px(ours)	1×	35.0	53.1	36.8	16.6	40.8	49.7
		△	(+3.2)	(+4.0)	(+3.2)	(+3.9)	(+4.2)	(+2.3)
		3×	35.7	54.0	37.4	16.3	41.3	50.8
		△	(+1.0)	(+1.5)	(+0.8)	(+1.8)	(+1.8)	(+0.2)
FasterRCNN-R50 Ren et al. (2015)	800px	3×	40.3	61.0	44.0	24.0	44.1	51.4
	600px	1×	36.8	57.5	39.8	20.6	40.5	49.2
		3×	39.3	59.9	42.4	22.6	42.8	52.7
	400px	1×	34.4	53.8	37.0	15.9	38.4	49.1
		3×	36.5	55.9	39.0	17.5	40.5	52.2
	600px(ours)	1×	39.5	59.9	42.9	22.9	43.3	52.2
		△	(+2.7)	(+2.4)	(+3.1)	(+2.3)	(+2.8)	(+3.0)
		3×	39.8	60.1	43.3	23.2	43.6	53.1
		△	(+0.5)	(+0.2)	(+0.9)	(+0.6)	(+0.8)	(+0.4)
	400px(ours)	1×	36.9	56.5	39.7	18.3	40.9	51.9
		△	(+2.5)	(+2.7)	(+2.7)	(+2.4)	(+2.5)	(+2.8)
		3×	37.6	56.9	40.7	18.8	41.9	52.7
		△	(+1.1)	(+1.0)	(+1.7)	(+1.3)	(+1.4)	(+0.5)

naNet Lin et al. (2017b) and Faster R-CNN Ren et al. (2015)[3], then re-compare LRDet with the low-resolution baselines in Table 3. Results show that with a 1× schedule, LRDet can achieve greater improvements in the multi-scale setting compared to the single-scale setting (+3.1 mAP *vs.*+2.4 mAP for RetinaNet and +2.7 mAP *vs.*+2.1 mAP for Faster R-CNN). Considering that detectors need more training iterations to converge when training with stronger augmentations, we also provide the results with a 3× schedule in Table 3. It is nontrivial to get improvement with 3x multi-scale training schedule especially in our self-distillation setting, but LRDet still deliver non-negligible gains. The above results prove the effectiveness and the robustness of LRDet across various detectors

[3]We do not provide multi-scale trained FCOS results in Table 3. As in MMdetection, the standard multi-scale setting for FCOS is not a 3× schedule. Instead, we give the multi-scale trained FCOS results in Table 6 and compare our LRDet with MSAD.

Table 4: Ablation study on LRDet with RetinaNet Lin et al. (2017b) on MS COCO.

(a) Component ablation.

Bridge	Inherit	AP	AP_S	AP_M	AP_L
–	–	35.4	18.0	39.5	48.3
✓	–	37.1	20.1	40.8	50.1
–	✓	37.0	19.9	40.8	50.0
✓	✓	**37.8**	**20.2**	**41.8**	**51.1**

(b) Stronger teacher networks in single-scale setting.

LR student	HR teacher	AP	AP_S	AP_M	AP_L
RetinaNet-R50	–	35.4	18.0	39.5	48.3
RetinaNet-R50 (ours)	RetinaNet-R50	37.8	20.2	41.8	51.1
	RetinaNet-R101	38.9	21.1	43.0	53.3
	RetinaNet-X101	39.7	21.2	43.7	55.1

(c) Stronger teacher networks with multi-scale 3x training schedule.

LR student	HR teacher	AP	AP_S	AP_M	AP_L
RetinaNet-R50 -600px	–	38.3	21.0	42.6	51.0
RetinaNet-R50 -600px (ours)	RetinaNet-R50	39.1	22.6	43.2	52.0
	RetinaNet-R101	40.5	22.7	44.5	55.4
RetinaNet-R50 -400px	–	34.7	14.5	39.5	50.6
RetinaNet-R50 -400px (ours)	RetinaNet-R50	35.7	16.3	41.3	51.8
	RetinaNet-R101	37.1	17.8	42.9	54.0

(d) Lightweight student networks. Our designs works for lightweight networks.

Detector	Resolution	AP	AP_S	AP_M	AP_L
RetinaNet -R18	600px	31.0	14.8	33.8	43.0
	400px	27.3	9.9	29.9	42.4
	600px(ours)	34.6	16.9	37.5	48.8
	400px(ours)	30.2	11.8	32.7	46.1
RetinaNet -Mbv2	600px	29.1	14.1	31.6	41.3
	400px	26.0	9.4	28.6	40.2
	600px(ours)	32.7	16.5	35.5	46.2
	400px(ours)	28.5	10.4	31.2	44.0

(e) Heterogeneous architectures.

LR student	HR teacher	AP	AP_S	AP_M	AP_L
RetinaNet-R50	–	35.4	18.0	39.5	48.3
RetinaNet-R50 (ours)	FasterRCNN-R50	37.0	20.0	40.8	50.8
	MaskRCNN-R50	37.4	20.6	41.4	50.8

with different settings. It should be noted that if not limited to self-distillation, our LRDet can get more considerable gains with 3x multi-scale training schedule (See Table 4c).

4.2 Ablation Study

We aim to find out how LRDet works in this section. We run a number of ablation studies with RetinaNet Lin et al. (2017b) using the single-scale setting. Note that except for the experiments in Table 4b and Table 4d, all other ones are with ResNet-50 He et al. (2016). Details are as follows.

Component ablation. LRDet transfers high-resolution knowledge from the high-resolution teacher to the low-resolution student via the bridge branch and weight inheriting. We ablate these two component and the results are shown in Table 4a. Distillation with our designed bridge branch helps the low-resolution student improve by 1.7 mAP, while the conventional framework can only improve the student by a maximum of 0.9 mAP (See Table 1). We also find that only using weight inheriting can also bring a 1.6 mAP improvement. And combination can achieve further improvement.

Stronger teacher models. The main purpose of our LRDet is to improve low-resolution detection with the help of a high-resolution teacher. In previous experiments, we constrain the variable factor to resolution only and adopt the same networks for both the teacher and the student. But LRDet is not limited to this self-distillation setting. Intuitively, stronger teachers typically lead to greater gains. In Table 4b and Table 4c, we explore teacher models with stronger backbones, such as ResNet-101 He et al. (2016) and ResNeXt-101 Xie et al. (2017), while maintaining ResNet-50 He et al. (2016) for the student model.

As shown in the Table 4b, LRDet enables the low-resolution student on 600×1000 to get further improvements in the single-scale setting by using stronger high-resolution teachers. Notably, when using RetinaNet-X101, the low-resolution detector achieves 39.7 mAP. It exceeds the low-resolution student baseline by 4.3 mAP, showing the great potential of our LRDet in improving low-resolution detection. Table 4c shows that adopting a stronger teacher can lead to considerable improvement even in the multi-scale $3\times$ training setting. Specifically, the LR student models on 600×1000 and 400×667 achieve 40.5 and 37.1 mAP, respectively.

Lightweight student networks. In addition to adopting stronger teacher models, we investigate the opposite direction: apply lightweight backbones for student models. We take ResNet-18 He et al. (2016) and MobileNet-V2 Sandler et al. (2018) as backbones for low-resolution students. To show better results, according to Table 4b, we use a strong high-resolution model (with ResNeXt-101 Xie

Table 5: Discussion on the choice of the high-resolution teacher with our LRDet. Results with different high-resolution teachers are provided.

LR	HR	AP	AP_S	AP_M	AP_L
400px	–	32.1	12.3	36.4	48.3
400px (ours)	800px	33.9	14.2	39.1	49.9
	600px	34.2	14.4	38.4	50.0
	800px & 600px	34.7	14.5	39.8	50.5

Table 6: Discussion on the format of the bridge detector. We conduct experiments on FCOS Tian et al. (2019b) with ResNet-50 and make a fair comparison at comparable FLOPs.

Method	Resolution	FLOPs	AP	AP_S	AP_M	AP_L
MSAD	400px	144.8G	39.7	21.7	42.9	55.0
LRDet	400px	**52.1G**	38.2	21.0	42.3	50.9
LRDet	600px	114.3G	**41.3**	**24.2**	**45.3**	**54.1**

Table 7: Generalization of LRDet. We extend our LRDet to instance segmentation models and human keypoint detection models

Task	Detector	Resolution	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_S	AP^{mask}_M	AP^{mask}_L
Instance Segmentation	Mask-RCNN -R50	800px	35.4	56.4	37.9	19.1	38.6	48.4
		600px	33.6	54.1	35.8	13.9	36.2	51.0
		400px	31.1	50.7	32.8	10.0	33.2	50.4
		600px(ours)	35.5	56.5	37.7	16.1	38.5	53.7
		△	(+1.9)	(+2.4)	(+1.9)	(+2.2)	(+2.3)	(+2.7)
		400px(ours)	32.5	52.2	34.5	11.4	35.2	52.5
		△	(+1.4)	(+1.5)	(+1.7)	(+1.4)	(+2.0)	(+2.1)
	SOLOv1 -R50	800px	33.1	53.5	35.0	12.2	36.1	50.8
		600px	31.9	51.4	33.5	10.5	34.7	50.3
		400px	28.8	47.3	30.2	6.5	30.5	50.8
		600px(ours)	33.5	53.9	35.6	11.6	37.0	52.5
		△	(+1.6)	(+2.5)	(+2.1)	(+1.1)	(+2.3)	(+2.2)
		400px(ours)	30.1	49.2	31.6	7.4	32.7	52.0
		△	(+1.3)	(+1.9)	(+1.4)	(+0.9)	(+2.2)	(+1.2)
Task	Detector	Resolution	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
Keypoint Detection	Swin-base	384×288	75.1	90.6	82.2	80.5	94.3	86.7
		256×192	73.1	90.0	80.9	78.8	94.1	85.7
		256×192(ours)	74.1	90.7	81.9	79.6	94.5	86.7
		△	(+1.0)	(+0.7)	(+1.0)	(+0.8)	(+0.4)	(+1.0)

et al. (2017) as the backbone) as the teacher. The results are shown in Table 4d. The effectiveness of our LRDet holds with lightweight models. Specifically, at 600 resolution, both RetinaNet-R18 and RetinaNet-mbv2 achieves a $+3.6$ mAP improvement.

Heterogeneous architectures. In previous experiments, our LRDet concentrate on homogeneous teacher-student detectors. However, the student detector for deployment is often significantly different from the teacher. Thus, we investigate the cross-resolution distillation among heterogeneous teacher-student pairs for a wide application. We use the two-stage detector (e.g., Faster-RCNN) to distill the single-stage detector (RetinaNet), and the experimental results are shown in the Table 4e. Note that the inherit strategy is not available when distilling between heterogeneous detectors. Our method deliver non-trivial gains with the heterogeneous teachers, which shows the potential for practical applications of LRDet.

4.3 More Discussions

In this section, we provide two further discussions on our LRDet. One is how to choose the high-resolution teacher for the low-resolution student. Another discusses the way to construct a bridge for low-resolution detection. Experiments and results are presented below.

Discussion on choosing high-resolution teachers. In our previous attempts, we adopt the 800×1333 high-resolution teacher for all low-resolution models. While in real applications, there may be different resolutions that can serve as high resolutions. This section discusses how to choose a high-resolution teacher for a low-resolution student. We conduct experiments by adopting 400×667 as the low-resolution and considering 600×1000 and 800×1333 as the high-resolution. Table 5 provide the low-resolution detection results with RetinaNet-R50 using different high-resolution teachers. Results show that the 600×1000 teacher brings more gains than the 800×1333 teacher. We conjecture that the resolution gap between the teacher and the student affects the knowledge transfer. Thus, choosing a high-resolution teacher with a relatively small resolution gap may bring more improvements with LRDet. Moreover, in order to further improve the student model, we use 600×1000 and 800×1333 teachers to distill student simultaneously, which yields a better result.

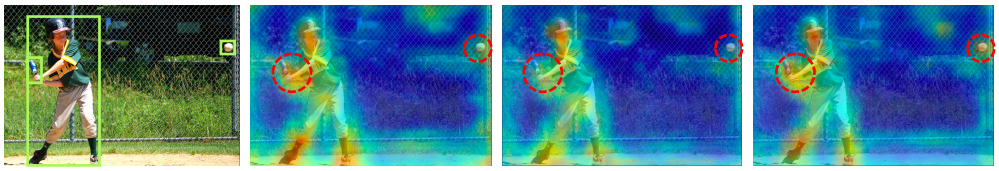

Figure 2: Visualization of FPN feature maps. LRDet helps the LR student to recover the attention on the medium and small objects.

Discussion on constructing the bridge. We try to discuss how to construct a bridge for low-resolution detection. As presented in our related works, MSAD Qi et al. (2021) also try to improve the results of low-resolution students by knowledge distillation. They apply a high-resolution teacher whose resolution is $2\times$ larger than the low-resolution. Thus, they propose to fuse teacher features and student features in a cross-level manner, which serve as a bridge to distill the low-resolution student. However, their bridge has two limitations. One is that it requires a $2\times$ differences to perform cross-level feature fusion. The other is that their LR student models use P2$\sim$P6 features, which is two times larger than the standard P3$\sim$P7. It is worth mentioning that using P2$\sim$P6 features naturally leads to a $\sim$1.5 mAP improvement, but it significantly increases computational cost. LRDet addresses their limitations. Our framework has no constraint on resolution and can be used for any two resolutions. Meanwhile, we use standard P3$\sim$P7 features for detection without any additional calculation during inference. We conduct experiments with FCOS to prove that LRDet can works better than MSAD[4]. Table 6 shows the comparison between LRDet and MSAD. The inference FLOPs of MSAD is much higher than ours when using the same low-resolution. To be fair, we compare the accuracy under the comparable FLOPs. The results show that LRDet outperforms MSAD while requiring less computation cost.

4.4 Generalization

To validate the generalization of our framework, we further conduct experiments on the instance segmentation task and human keypoint detection task on the MS COCO benchmark Lin et al. (2014). For instance segmentation task, we experiment on Mask-RCNN He et al. (2017) and SOLO-v1 Wang et al. (2020a) in the self-distillation settting. For human keypoint detection task, Swin-transformer Liu et al. (2021) is adopted. Table 7 shows that our method improves the LR students consistently, indicating that LRDet has great potential to generalize to various position-sensitive tasks.

4.5 Visualization

By visualizing the feature maps, we verify that LRDet trains stronger low-resolution detector. As shown in Figure 2, the LR model pays less attention on objects than the HR model, especially on medium and small ones. And our LRDet successfully helps the LR student to recover the attention on these medium and small objects, which further helps the student to be more accurate.

5 Conclusion

In this paper, we focus on improving low-resolution detection with high-resolution teachers. We point out that the resolution gap and supervision mismatch between HR teachers and LR students hinder the conventional distillation framework. Then we propose LRDet as an simple yet effective framework to perform distillation across resolutions. The key insight of LRDet is the introduced bridge branch, which can resolve the above problems caused by different resolutions. With the bridge branch, the superior knowledge from the HR teacher can be smoothly transferred to the LR student. We apply LRDet to various detectors on MS COCO benchmark using different image resolutions. Extensive experimental results demonstrate the effectiveness of our LRDet. Moreover, we conduct extensive ablation studies and discussions on LRDet to provide a better understanding.

[4]All experiments in Table 6 use the high-resolution model provided by MSAD as the HR teacher and adopt $1\times$ training schedule.

REFERENCES

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851, 2021.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2154–2164, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1921–1930, 2019.

Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34, 2021.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 6356–6364, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14443–14453, 2021.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5311–5320, 2021.

Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3578–3587, 2018.

Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. *Advances in neural information processing systems*, 31, 2018.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019a.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019b.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019.

Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pp. 649–665. Springer, 2020a.

Yikai Wang, Fuchun Sun, Duo Li, and Anbang Yao. Resolution switchable networks for runtime efficient image recognition. In *European Conference on Computer Vision*, pp. 533–549. Springer, 2020b.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9657–9666, 2019.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4643–4652, 2022.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020.

A PSEUDO CODE

Algorithm 1 LRDet: PyTorch-like Pseudocode

```
# f_s, f_t: Backbone of student and teacher
# fpn_s, fpn_t: FPN of student and teacher
# h_s: Head(s) of student
# alpha: Hyperparameter to balance loss

# load low-resolution img and gt
for img, gt in loader:
    # get high-resolution img and gt
    hr_img, hr_gt = upsample(img, gt)

    # lr student process
    s_feat = f_s(img)
    s_fpn_feat = fpn_s(s_feat)
    s_detection = h_s(fpn_s_feat)

    # bridge branch process
    b_feat = upsample(s_feat)
    b_fpn_feat = fpn_s(b_feat)
    b_detection = h_s(fpn_b_feat)

    # hr teacher process
    t_feat = f_t(hr_img)
    t_fpn_feat = fpn_t(t_feat)

    # standard detection loss
    s_loss = det_loss(s_detection, gt)
    b_loss = det_loss(b_detection, hr_gt)
    # feature distillation loss
    d_loss = distill(b_fpn_feat, t_fpn_feat)

    # update student
    loss = s_loss + b_loss + alpha * d_loss
    loss.backward()
    update(f_s, fpn_s, h_s)
```

LRDet can be implemented with several simple changes to the conventional distillation framework. The pseudo-code is provided in Algorithm 1. The high-resolution images and ground truth boxes are obtained by upsampling the low-resolution ones. We adopt standard detection loss to supervise the student model and the bridge branch. Meanwhile, we use the simple MSE as the distillation loss. It should be noted that our framework is compatible with many feature distillation algorithms, which are discussed next (see Table 8).

B MORE ABLATION STUDIES

Table 8: Robustness evaluation of LRDet on various types of distillation losses with RetinaNet-Res50 Lin et al. (2017b) on MS COCO Lin et al. (2014).

Loss	AP	AP_S	AP_M	AP_L
MSE	37.8	20.2	41.8	51.1
L1	37.7	20.1	41.6	51.3
CWD Shu et al. (2021)	37.7	20.3	41.4	51.6
FGD Yang et al. (2022)	37.8	20.1	41.8	51.0

Our LRDet is a simple framework to improve the low-resolution detection, which *does not require the sophisticated design of the model structure and careful parameter tuning.*

Table 9: Study on the sensitivity of α. Our LRDet is robust to the loss weight α in Eq. (1) in the main paper.

Reduction	α	AP	AP_S	AP_M	AP_L
mean	1.0	37.7	20.0	41.2	51.9
sum	0.0005	37.8	20.1	41.2	51.7
	0.001	37.8	20.4	41.4	51.6
	0.005	37.8	20.2	41.8	51.1

Robustness evaluation of LRDet on loss types. In Table. 8, we evaluate the robustness of our method on various distillation losses. We conduct experiments on the single-scale setting, *i.e.*, RetinaNet-Res50-800px Lin et al. (2017b) as HR teacher and RetinaNet-Res50-600px as LR student. We use four-type distillation losses, including MSE, L1, CWD Shu et al. (2021), FGD Yang et al. (2022). As shown in Table 8, different losses all lead to significant improvements. Compared to the well-designed distillation losses such as FGD and CWD, the simple MSE or L1 loss can already achieve comparable performance, indicating that our framework is robust to the type of distillation loss.

Study on the sensitivity of α. We further evaluate the robustness of LRDet on the loss weight α in Eq. (1) in the main paper. In Table 9, we adjust the loss reduction format and the loss weight of the MSE loss. Other settings follow Table 8. As shown in Table 9, a simple *mean* reduction leads to a decent improvement, while using *sum* reduction achieves better results. Meanwhile, different loss weights all lead to considerable improvements, showing the robustness of LRDet on the loss weight[5].

C More Visualizations

In the main paper, we have visualized the FPN feature maps, here we report more visualizations of the detection results to show the effect of our LRDet. As shown in Figure 3, LRDet improves object detection on low-resolution images, especially the detection of small and medium-sized objects in crowded and overlapping scenes.

[5]We note that the *sum* reduction brings slightly higher gains for small target detection. We thus adopt the *sum* reduction by default.

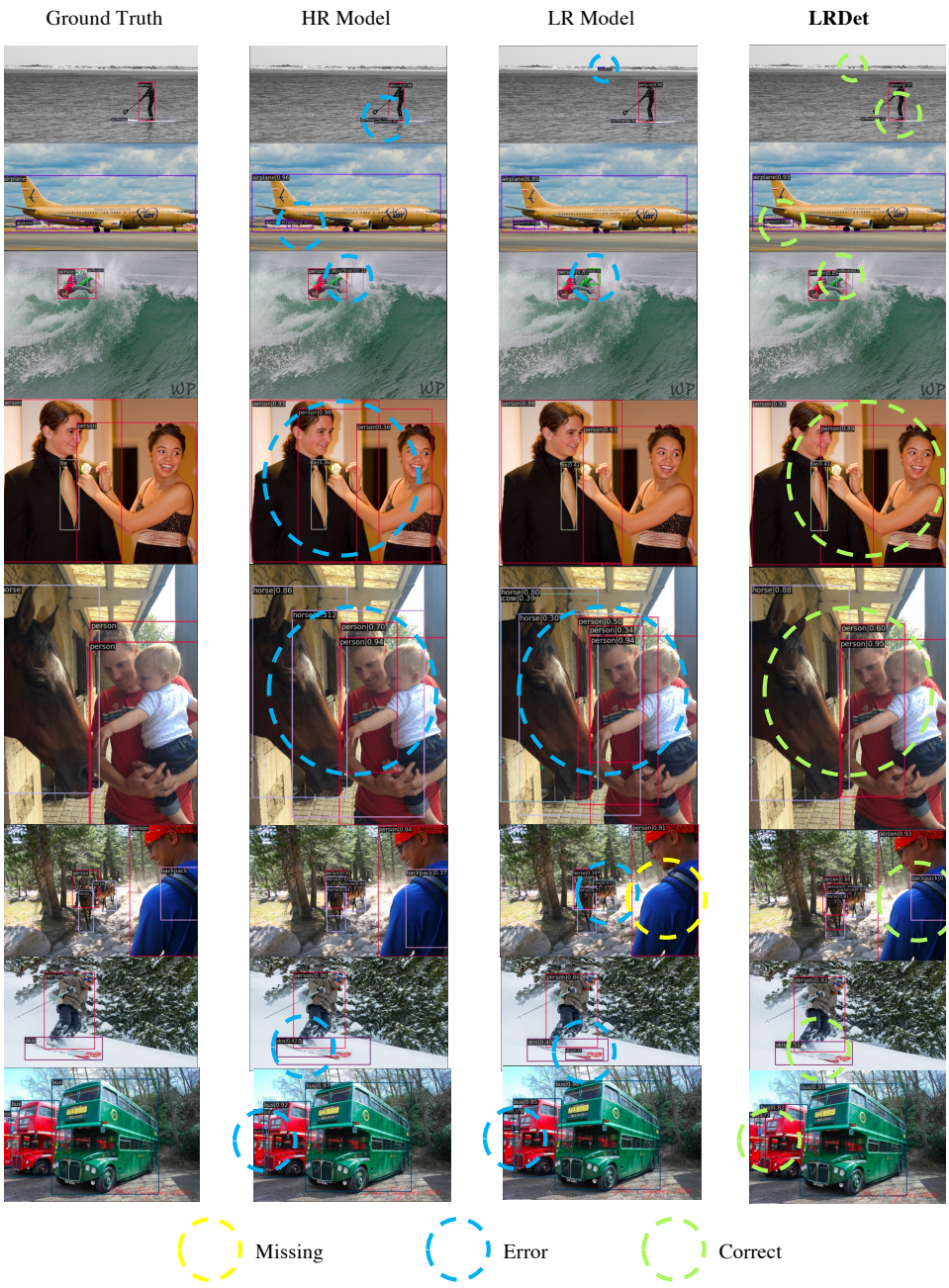

Figure 3: **Visualization of detection results.** Missing: false negative bounding boxes. Error: false positive bounding boxes. (Best viewed in color and magnification.)